2009

Smart Templates for peak pattern matching with comprehensive two-dimensional liquid chromatography

Stephen E. Reichenbach
University of Nebraska - Lincoln, reich@cse.unl.edu

Peter W. Carr
University of Minnesota - Duluth

Dwight R. Stoll
University of Minnesota

Qingping Tao
GC Image, LLC, Lincoln, NE, qtao@gcimage.com

Presented to the 32nd International Symposium on High Performance Liquid Phase Separations and Related Techniques;
published in *Journal of Chromatography A* 1216:16 (April 17, 2009), pp. 3458–3466; doi: 10.1016/j.chroma.2008.09.058

Published online September 21, 2008.

Smart Templates for peak pattern matching with comprehensive two-dimensional liquid chromatography

Stephen E. Reichenbach,[1] Peter W. Carr,[2] Dwight R. Stoll,[2] and Qingping Tao [3]

1. University of Nebraska–Lincoln, Computer Science and Engineering Department,
Lincoln, NE 68588-0115, USA. *Corresponding author* — tel 402 472-5007, fax 402 472-7767, email reich@cse.unl.edu

2. University of Minnesota, Department of Chemistry, Minneapolis, MN 55455-0431, USA

3. GC Image, LLC, P.O. Box 57403, Lincoln, NE 68505-7403, USA; http://www.lcxlc.com ; email qtao@gcimage.com

Abstract
Comprehensive two-dimensional liquid chromatography (LC × LC) generates information-rich but complex peak patterns that require automated processing for rapid chemical identification and classification. This paper describes a powerful approach and specific methods for peak pattern matching to identify and classify constituent peaks in data from LC × LC and other multidimensional chemical separations. The approach records a prototypical pattern of peaks with retention times and associated metadata, such as chemical identities and classes, in a *template*. Then, the template pattern is matched to the detected peaks in subsequent data and the metadata are copied from the template to identify and classify the matched peaks. *Smart Templates* employ rule-based constraints (e.g., multispectral matching) to increase matching accuracy. Experimental results demonstrate Smart Templates, with the combination of retention-time pattern matching and multispectral constraints, are accurate and robust with respect to changes in peak patterns associated with variable chromatographic conditions.

Keywords: two-dimensional chromatography, liquid chromatography, chemical identification and classification, pattern matching, pattern recognition

1. Introduction

Comprehensive two-dimensional liquid chromatography (LC × LC) provides an order-of-magnitude increase in peak separation capacity over one-dimensional high-performance liquid chromatography (HPLC) [1]. With its greater separation power, LC × LC reduces co-elutions, which reveals otherwise unseen chemical complexity and allows improved quantitation, and exposes multidimensional structure–retention relationships, which can be exploited for improved chemical identification and classification. Since early work on LC × LC [2, 3], research and development have significantly improved and refined LC × LC technologies [4–6]. The future for LC × LC is especially promising for important but challenging biochemical applications [7], including proteomics [8, 9] and metabolomics [10, 11], which typically contain thousands of constituents with widely varying concentrations within the same sample.

Although LC × LC holds great promise, the lack of software for data processing and automated analysis is a major obstacle to its effective widespread application. In a recent survey of fast LC × LC, Stoll et al. concluded that "the paucity of efficient, convenient and sufficiently powerful data analysis tools" is "the greatest impediment to wide application of 2DLC." [5, p. 39] Guiochon et al. write: "More sophisticated problems need to be solved. They deal with how to help analysts in making sense of these large data arrays, in using these painfully acquired data to solve important analytical problems, in how actually to handle these data and turn them into relevant numbers."[6, p. 159]

The need for more rapid and effective analytical software is especially critical for biological separations:

- "The need for computational methods is evident in order to find peaks that correlate with phenotypes and, equally importantly, in order to assess their statistical significance." [12, p. 2]
- "The lack of effective generic procedures for routinely detecting differences in global protein patterns across many different samples hinders the discovery of new biomarkers." [12, p. 984]
- "Improvements/development of bioinformatics packages are urgently needed for the conduction of all steps of proteomic studies." [14, p. 17]
- "[T]he primary bottleneck in high throughput proteomic production 'pipelines' is in many cases no longer the rate at which the instrument can generate data, but rather it is in quality analysis and interpretation of the results to generate confident protein assignments." [15, p. 497]

Because of the size and complexity of LC × LC data, the lack of software is even more acute than for some other analytical technologies and is one of the most significant impediments to the adoption of LC × LC. This problem is evident in many recent publications of researchers pioneering LC × LC. As Dixon et al. note in reviewing LC × LC for biomedical and pharmaceutical analysis, data processing and analysis for biological separations is already difficult but will be even more so now that "*n*-dimensional data acquisition is a reality" [16, p. 526].

LC × LC offers increased information capacity for complex chemical separations, but with its greatly increased performance, LC × LC generates data in significantly larger quantity and with significantly greater complexity than one-dimensional HPLC. Compared to data from one-dimensional HPLC, LC × LC data has many times more data points, an order-of-magnitude greater peak capacity, and added data dimensionality. Analysis of LC × LC data is challenging and requires computer automation and assistance. LC × LC transforms chemical samples into raw data; information technologies are required to transform LC × LC data into useful information.

This paper addresses the problem of automatically identifying and classifying the peaks of interest in chromatograms of similar mixtures with possibly variable chromatographic conditions. A popular method for peak identification in one-dimensional chromatography is to define retention-time windows for the peaks of target compounds. Under repeatable, reproducible, and tightly controlled chromatographic conditions, the peaks for target compounds will fall reliably within fixed retention-time windows. However, narrow windows may be required for peaks with nearby neighboring peaks (to avoid false identifications) and, with narrow windows, even slightly different chromatographic conditions may cause a peak to drift outside its window. Here, "drift" is used to characterize a local variation which may be related to more complex systemic variations as might be caused by stationary phase aging due to instability or build-up of contaminants, instrument aging, lack of sufficient temperature control, and variations in pumping system performance. The problems related to retention-time drift in peak identification for LC × LC are more complex than for one-dimensional HPLC.

This paper describes a powerful approach and specific methods for peak pattern matching to identify and classify constituent peaks in data from LC × LC and other multidimensional chemical separations. The approach records a prototypical pattern of peaks with retention times and associated metadata, such as chemical identities and classes, in a *template*. Then, the template pattern is matched to the detected peaks in subsequent chromatograms and the metadata are copied from the template to identify and classify the matched peaks. *Smart Templates* employ rule-based constraints (e.g., multispectral matching) to increase matching accuracy. For example, the Smart Template may record the expected spectrum of a target compound and then require that a matched chromatographic peak have a sufficiently similar spectrum. The constraints in Smart Templates may be written by hand, based on expert knowledge, or constructed automatically. Experimental results demonstrate that the method is accurate and robust with respect to changes in peak patterns due to variations in chromatographic conditions.

Section 2 outlines the chromatographic acquisition of the experimental data on which the methods are demonstrated. Section 3 develops an algorithm for two-dimensional gradient background detection, modeling, and removal. Background removal is a much more serious issue for LC × LC than comprehensive two-dimensional gas chromatography (GC × GC) due to the large signals generated by changes in eluent composition during gradient elution. The algorithm modifies a method developed for GC × GC to account for the dynamic response in the second-column gradient separation of LC × LC, thereby allowing accurate peak detection and quantification. Section 4 presents simple methods for two-dimensional peak detection and multispectral matching for chemical identification. Section 5 details the use of templates and template matching for recognizing patterns of peaks in LC × LC data. Section 6 describes how Smart Templates with rule-based constraints can significantly improve template matching accuracy and describes how constraint rules can be constructed automatically. Section 7 contains concluding remarks about the applicability of the approach to other types of detectors and other types of multidimensional chemical separations.

2. Data acquisition

The example data analyzed in this paper were acquired at the University of Minnesota in a series of 64 injections of: (a) water (four injections near the end of the series); (b) a standards mixture with potassium nitrate, tryptophan, hydroxytryptophan, indole-3-acetic acid, indole-3-propionic acid, indole-3-acetonitrile, and tyrosine (6 injections interspersed in the series); (c) a control urine sample (14 injections interspersed in the series); and (d) experimental urine samples (40 injections, four of which failed). For the urine analyses, a 460 µL aliquot of each urine sample was transferred to a HPLC vial. To each vial, 40 µL of 70% perchloric acid was added to precipitate proteins and this solution was allowed to stand for 10 min, followed by filtration with a small 0.2 µm PTFE syringe filter. The filtrate was collected in a new HPLC vial to which 55 µL of 10 M potassium hydroxide was added. This solution was centrifuged for 5 min to pellet the solid potassium perchlorate. For the experimental samples, the resulting solution was either diluted 9:10, 1:4, or 1:16 using 20 mM sodium phosphate, 0.1 mM EDTA, pH 6. Then, the samples were injected without further treatment.

In the dual gradient-elution system developed by Stoll et al., pictured in Figure 1, the first column is comprised of a conventional gradient-elution HPLC system and reversed-phase LC column [17]. The effluent from the first column is captured alternately in Loop 1 or Loop 2 (denoted L1 and L2 in Figure 1) of the 10-port valve shown in the center of the figure. The stored effluent is injected into Column 2, the second dimension of the separation, and subjected to gradient elution by the dual gradient pumping system (Pumps B and C). The very rapid second separation uses a very short narrow column with high temperature (> 100°C) and high flow rate (3 cc/min) to achieve very

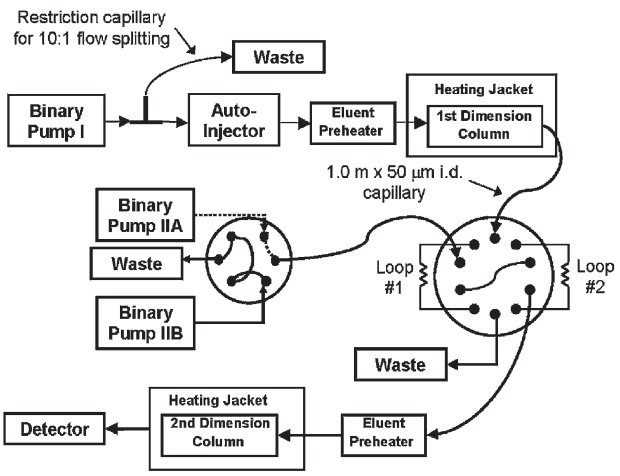

Figure 1. Instrumentation for comprehensive two-dimensional liquid chromatography (LC × LC) [17].

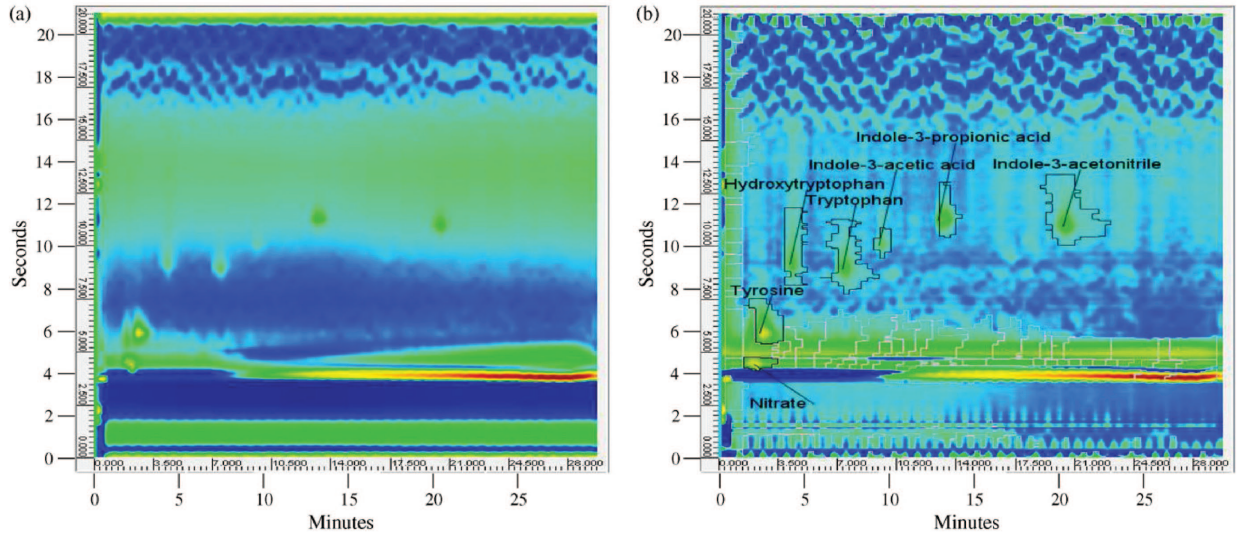

Figure 2. (a) An image of LC × LC data for the standards mixture. Before background correction, the dynamic range of the background obscures peaks. (This and other data visualizations were rendered with GC Image®software for comprehensive two-dimensional chromatography [18].) (b) After correction, the background values in the broad center of the image are near zero. The detected peaks are much clearer and the peaks of interest are outlined in black with black labels.

high linear velocity, allowing these separations to complete within 21 s. This is extraordinarily fast for liquid chromatography and the resulting peaks are very narrow (<0.5 s half-height width). The two independent pumps and valve allow switching between the two systems to minimize the effect of gradient dwell volumes. Otherwise, the chromatography would be slowed substantially and the retention-time reproducibility in the second dimension would be greatly compromised.

Although gradient elution in the second dimension is not as simple as isocratic elution, it is essential for three reasons. First, gradient elution gives higher peak capacity than isocratic elution. Second, a strong final eluent insures that everything elutes before the next separation starts. Third, gradient elution allows the diluted sample from the first dimension to be focused at the top of the second column, thereby improving the second dimension peak width when the first dimension system is delivering the analytes in strong eluent.

In these runs, the gradient in the first column runs from 0 to 23 min, returns to the initial composition at 23.01 min, and is held there until the end of the cycle (29.75 min). The first-column dead-time is 1.0 min. The gradient in the second column runs from 0 to 18 s, returns to the initial composition at 18.6 s, and is held there until the the end of the cycle (21 s). The second-column dead-time is 1.3 s.

The data was collected with a PhotoDiode Array Detector (DAD) over the wavelength range 200–700 nm sampled in 4 nm intervals at 40 Hz for 29.75 min and written to a file by Agilent ChemStation software. The data for each run contained 71,400 data points, each with 126 spectral intensities, for a total of nearly 9 million intensities per run. As described in the following sections, the data was read from the ChemStation UV file, restructured as a series of 85 secondary chromatograms, each 21 s long, and processed for background removal, peak detection, and peak identification with GC Image®LC × LC Software.

3. Preprocessing

Figure 2 a shows a pseudocolor image of one of six LC × LC chromatograms acquired for the standards mixture. The value of each pixel of the image is the total intensity count (TIC) of the ultraviolet (UV) spectral absorbance at the indicated first and second dimension retention times (respectively, the abscissa from left-to-right in minutes and the ordinate from bottom-to-top in seconds). (The UV TIC is computed as the sum of the responses, measured in milli-absorbance units (mAU), in all spectral channels, just as the total ion count is summed intensities for mass spectrometry.) The pixels are automatically pseudocolorized with Gradient-Based Value Mapping (GBVM) [19], which effectively uses the color scale to emphasize local differences in the data, even for variable data with a large dynamic range. (A small region containing the gradient front in the lower-right of the image are excluded from the GBVM computation.) The color map and the value map function are shown in Figure 3. For this data, the colorization shows each of the seven chemical peaks (discussed later) and the significant variations in the background values (discussed here). The background values, which can be seen directly wherever there is no chemical peak, vary greatly across the second column separations (bottom to top) and to a lesser

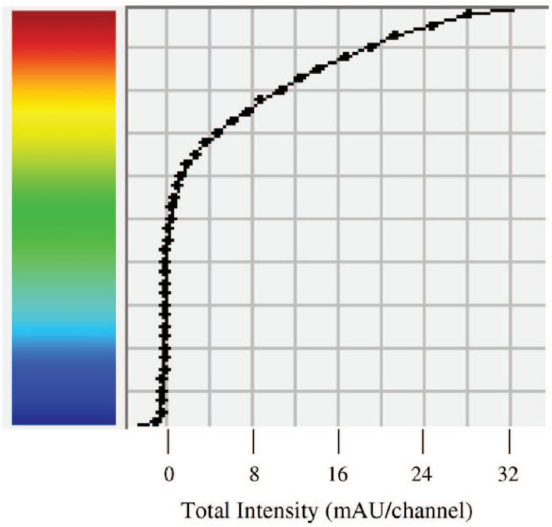

Figure 3. Color map and Gradient-Based Value Mapping (GBVM) function [19]. The function maps intensity values along the horizontal axis to a pseudocolor on the vertical axis.

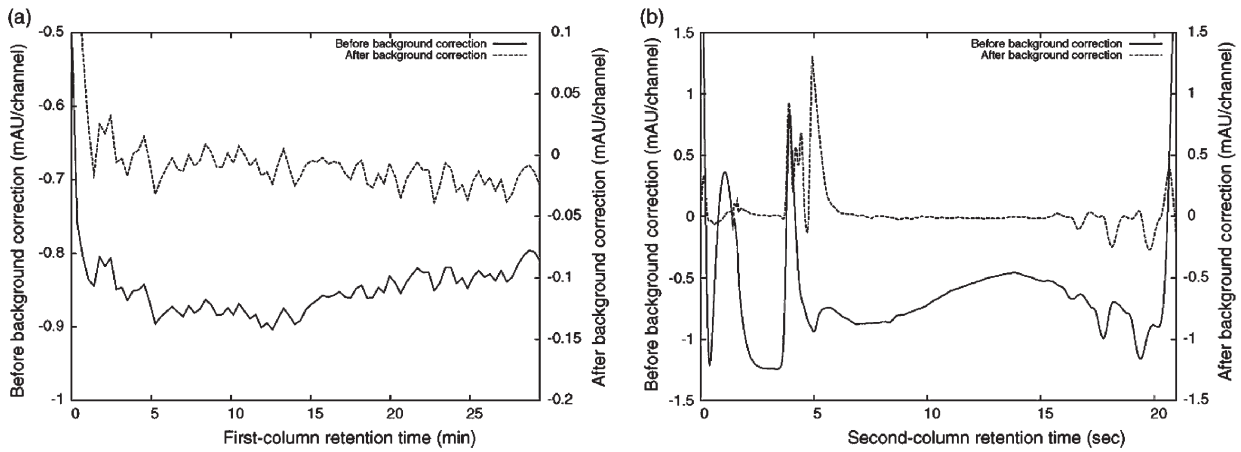

Figure 4. (a) Background values before (solid line) and after (dashed line) correction along a single row in the first dimension. A row with no analyte peaks was selected so that the values reflect only the baseline and noise. After correction, the values fluctuate in a small range centered very close to zero. (b) Background values before (solid line) and after (dashed line) correction along a single column in the second dimension. This secondary chromatogram with analyte no peaks was selected so that the values reflect only the baseline and noise. After correction, the values in the region of analysis are very close to zero.

extent across the first column separations (left to right). Note the increase in the middle of the second-column chromatograms between 7.5 and 13 s — from bottom-to-top, the color changes from blue to green — nearly obscures the peaks. The background values must be removed for accurate peak detection and quantification.

Background correction is performed with a new algorithm based on a method developed by Reichenbach et al. [20] for GC × GC. The GC × GC background correction method builds statistical models of the background values (by tracking neighborhoods around the smallest values as a function of time) and the noise (by parameterizing a Gaussian distribution for those neighborhoods) and then subtracts the background model from the data. That approach was modified in two important respects for LC × LC. First, because variations in the gradient separation background may be positive or negative, the LC × LC background correction algorithm must track the "middle" values (rather than the smallest values) by disregarding periods in which there are rapid changes or extreme values. Second, the background values vary greatly along the secondary separations, so the LC × LC background correction algorithm must model the background in both dimensions. With these important modifications, the LC × LC background correction algorithm is applied in each of the spectral channels.

The LC × LC background correction algorithm successfully corrects the background values in the regions of the chromatogram in which chemical analysis is performed. Figure 4 a graphs the values both before and after correction along the first dimension at a single row of data values (at 7.725 s of the second-column separations, left-to-right in Figure 2a). This row of data values was selected because no peak in any second-column separation is resolved at that time, so the values reflect only the baseline and noise. Before correction, the values decrease slightly from about −0.8 to −0.9 mAU (average-per-channel over all wavelengths) through the first half of the separation and then increase slightly to about −0.8 mAU at the end. (The spike at the initial sample falls outside the chromatographic range for chemical analysis and so is irrelevant.) After correction, the background values fall in a small range around zero (approximately −0.03 to 0.03 mAU), as desired. The local fluctuations related to noise remain, but the corrected baseline is very close to zero.

Figure 4 b graphs the background values along the second dimension (at 10.850 min of the first-column separation, bottom-to-top in Figure 2a). Before correction, the background values fluctuate significantly, especially at the beginning and end of each secondary separation (the bottom and top in Figure 2a). Some of the variations, such as those across the broad middle of the secondary separations are consistent across the image (left-to-right). Others, such as those at the top of the image are variable. In some regions outside the chromatographic range where chemical analysis is performed, the values change rapidly and inconsistently (e.g., the blotchy region at the top of the image) and are not fit by the smooth background model used by the algorithm. However, across the broad middle of the second column separations, the region in which chemical analysis is performed, the LC × LC background correction algorithm flattens the background values to near zero, as desired.

The resulting image of the data after background correction is shown in Figure 2 b. The background values across the center of the chromatogram are near zero and the chemical peaks (whose detection is described next) are clearer against the more uniform background. It is worth noting again that the colorization emphasizes the small variations in the background much more than would linear value mapping.

4. Peak detection and spectral identification

The chromatographic peaks are detected in two dimensions using the drain algorithm [21], a modified and inverted version of the watershed algorithm [22], on the LC × LC TIC. Multivariate chemometric methods for peak detection that aim to unmix or deconvolve co-eluted peaks based on differences in multispectral signatures (e.g., [23]) could detect more peaks, but those methods often are not robust enough for automation. Multivariate peak detection algorithms are an area of active research to address issues such as delineating regions for analysis (because many methods are not computational efficient enough to apply to all the data) and nonlinearity (e.g., peak shape changes related to column loading). Here, the drain algorithm works well enough for demonstrating the utility and power of Smart Templates for peak identification and classification.

The drain algorithm detects peaks from the top, down to the surrounding valleys, in two dimensions. With thresholds on the chromatographic footprint (i.e., the temporal area, which is the 2D analog of peak width) and apex value (the largest TIC among data points in the peak), the algorithm detects peaks

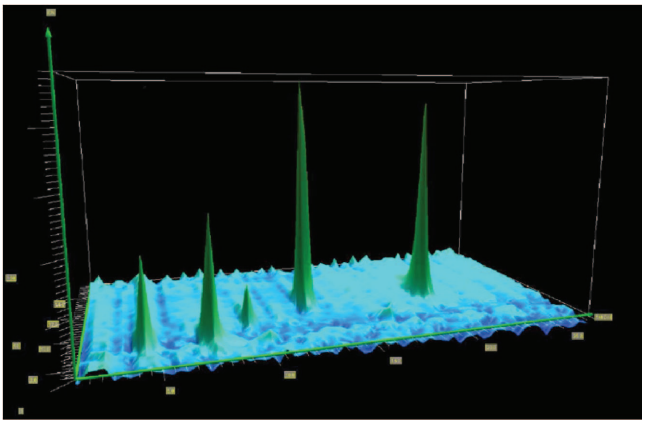

Figure 5. A three-dimensional perspective view of the center of the LC × LC image, with peaks for the five indoles in the standards mixture rising above the noise after background correction.

for each of the compounds in the standards mixture. In Figure 2 b, the footprints of the detected peaks are outlined, with the peaks of interest outlined in black. Other detected peaks, caused by artifacts and which are not in the region of analytical interest, are outlined in gray. The region with the five indoles in the mixture, which appear in the center of the image, is shown in three-dimensional perspective view in Figure 5. The linear vertical scale shows the extent to which the pseudocolor value mapping emphasizes the small variations in the background (while also clearly showing the peaks).

The spectra of the indole peaks in the image were compared to a database with the UV absorbance spectra of 26 indoles [24] using seven metrics (listed with the rate of correct identification for the five peaks in each of six images): Euclidean distance (70%), correlation (63%), first-derivative correlation (73%), absolute value difference(63%), first-derivative absolute value difference(67%), least squares (67%), and first-derivative least squares (73%). The database spectra were acquired with a different system at a different time and so tested the impact of reproducibility on multispectral identification. Each of the spectral matching metrics performed similarly well (63–73%).

For this sample mixture, chemical identification of the peaks by spectral matching is feasible: there are few peaks and the compounds in the mixture are known, so incorrect matches can be dealt with by a process of elimination from the list. In this example, ambiguous identifications for some peaks were established in this way. The rates of correct matches for each constituent compound across all metrics (seven metrics in each of six images) were: indole-3-acetonitrile (100%), indole-3-propionic acid (95%), hydroxytryptophan (67%), tryptophan (45%), and indole-3-acetic acid (33%).

In a complex mixture with many unknown compounds, UV detectors typically are not selective and sensitive enough for automatically identifying compounds with high confidence. Moreover, the multispectral matching typically requires human interaction to correct and validate the identifications, which is tedious and time-consuming for many chromatograms with many peaks. Smart Templates, described next, combine multispectral matching with chromatographic pattern recognition for more robust chemical identification, requiring far less human interaction to validate results and allowing full automation in some applications.

5. Templates and template matching

Template matching is based on the observation that the peaks in the two-dimensional retention-time plane form a pattern (or template) that can be recognized from one chromatogram to the next. Of course, this approach works only if the chemical compositions of the mixtures are similar so that the chromatograms exhibit many similar peaks. First, one or more chromatograms are carefully analyzed to identify peaks of interest and the pattern of those peaks, with their analyses, is recorded in a template. The analytical metadata (i.e., information about the peaks of interest, not including the intensity data itself) may include chemical identifications for some peaks, groupings of peaks (e.g., all peaks of a chemically related class), or even just the presence of a peak in the data (e.g., for comparisons between chromatograms to identify condition-related biomarkers). Next, given a new chromatogram, the unknown peaks can be identified by template matching. In template matching, the peaks in the template are matched to (paired with) detected peaks in the new chromato-

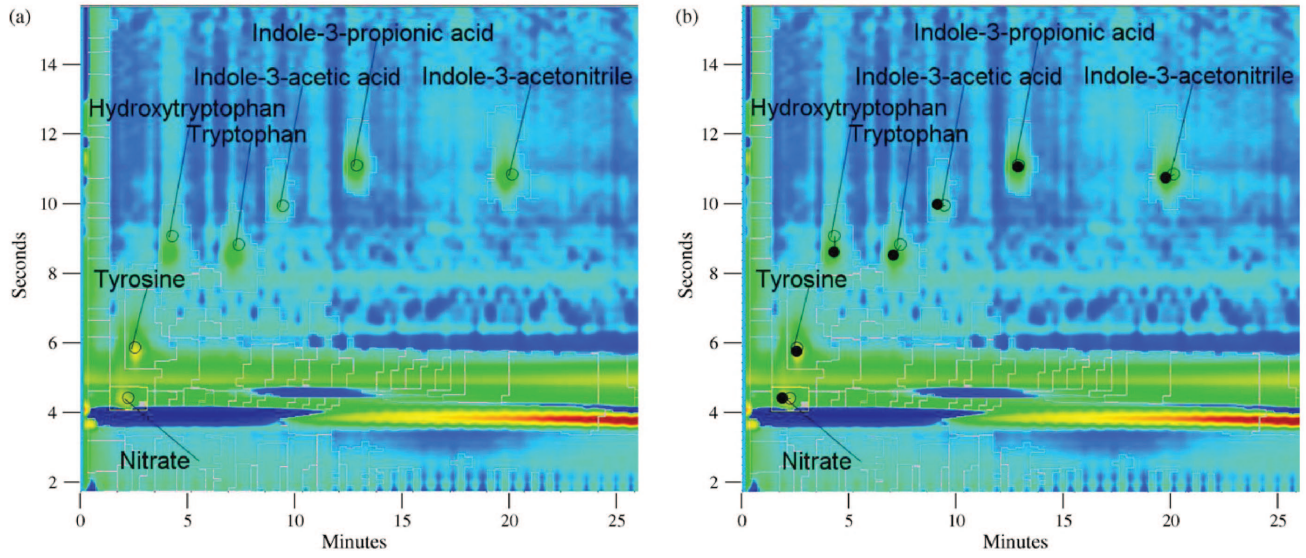

Figure 6. (a) The template from the first of six chromatograms of the standards mixture (with expected peak locations indicated with black open circles and labels) overlaid on the third of those chromatograms (with detected peaks outlined in gray). The alignment of the expected peak pattern to the detected peaks is close, but not perfect. (b) The matching of the template peaks to the detected peaks. The matched peaks are shown with filled circles with a connecting line to the corresponding template peak.

Table 1. Transformations for matching standards mixture templates and peak patterns

Template sequence #	Target sequence #	Translation (1)	Translation (2)	Scaling (1)	Scaling (2)
1	2	0.0000	−0.0711	1.0000	1.0119
2	20	−0.2493	−0.1014	0.9924	0.9788
20	38	−0.1069	0.0278	0.9990	1.0032
38	63	−0.2007	0.1883	0.9851	1.0286
63	64	0.0000	−0.0458	1.0000	1.0042
1	64	−0.5480	−0.0036	0.9771	1.0273

Translation units are the inter-sample times (21 s in the first dimension and 0.025 s in the second dimension). Scaling has no units of measure.

gram. Then, the analytical metadata (including peak identifications) are copied from peaks in the template into corresponding peaks in the new dataset.

This section begins with a simple example of a standards mixture with few peaks in order to illustrate templates and how template matching works and then proceeds to consider a more complex analysis. An example template and template matching are shown in Figure 6. Figure 6a shows the template peak pattern and metadata, with open black circles and labels, recorded from the first of the six chromatograms of the standards mixture (the chromatogram in Figure 2b), overlaid on an image of the third chromatogram of the standards mixture, with the detected peaks outlined in gray. As can be seen, the alignment of the template from the first standards chromatogram to the peaks of the third standards chromatogram is not perfect, but the template pattern matches the pattern of detected peaks well enough (i.e., within small retention-time windows) that correspondences can be established. Figure 6b shows the matches established for the example, with a filled black circle for each matched peak and a connecting line to the template peak with which it is matched. Then, the analytical metadata (here, the chemical identities of each peak) are copied from the template into the new dataset, thereby automatically identifying the peaks in the new chromatogram. In this way, peaks in the new chromatogram are identified by the metadata of their matching template peaks.

An important issue for template matching is retention-time "drift". Over the course of a long sequence of chromatographic runs, the pattern of the peaks may change, reflecting changes in the chromatographic conditions, such changes in the retentive properties of a column(s). Ni et al. [25] showed that GC × GC peak pattern variations over widely differing chromatographic conditions, such as temperature programming and pressure, can be modeled well by affine transformations. (Affine transformations are linear, geometric transformations, e.g., a sequence involving rotation, scaling, and translation/shifting.) Applying a geometric transformation (e.g., shifting/translating and scaling) to the template can bring its peak pattern into better alignment with the peaks of the new data so that peaks are matched more accurately. The template matching algorithm searches its transformation space for the model parameters that provide the best match—defined as allowing the most matches between template peaks and chromatographic peaks (within the allowed retention-time windows) [26]. The template matching algorithm used here [27] has a transformation model with translation and scaling in each of the two dimensions (parameterized by minimum and maximum translation and minimum and maximum scaling in each dimension) and a retention-time window (parameterized by width and height) within which the transformed template peaks may be matched to detected peaks after the template transformation. The approach allows for other transformation models, but this model has been validated for wide-ranging chromatographic variations [25] and has worked well in practice (e.g., in the examples shown here).

In the example of Figure 6 b, the matching algorithm finds a transformation with translation (−0.25 min, −0.17 s) and scaling (0.99, 0.99). With that transformation of the template, every matched chromatographic peak is within the specified retention-time window of the corresponding template peak. Other template components such as text labels, graphical objects such as polygons to delimit peak sets, and chemical symbols are geometrically transformed with the transformation established for the peak pattern.

Retention-time drift can be seen in the sequence of six chromatograms for the standards mixture, which were acquired within a longer sequence of 64 chromatograms. As shown in Table 1, the first of the standards runs was the first of the 64 runs, the second was the 2nd, the third was the 20th, and so on. (The first standards run was not the target of matching.) Table 1 presents the transformations for the matching of the peaks in the second standards run with the template from the first, for the matching of the peaks in the third standards run with the template from the second, and so on.

The table shows several notable trends. First, for the runs adjacent in the full sequence, standards runs one and two (runs one and two in the full sequence, the first row in Table 1) and standards runs 5 and 6 (runs 63 and 64 in the full sequence, the fifth row in Table 1), the matching transformation is very close to the identity transformation of translation (0,0) and scaling (1,1). Second, through the sequence, there is

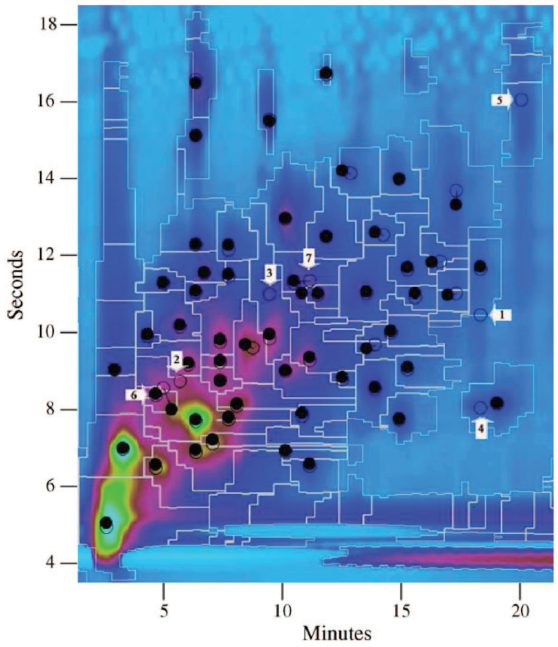

Figure 7. Template matching for a control urine sample chromatogram. Arrow 1: peak error, peak not detected cannot be matched. Arrow 2: peak error, merged peak not detected cannot be matched. Arrow 3: peak error, merged peak not detected cannot be matched. Arrow 4: peak error, merged peak not detected cannot be matched. Arrow 5: match error, peak too distant not matched. Arrow 6: match error, merged peak causes incorrect peak match. Arrow 7: match error, merged peak causes incorrect peak match.

a monotonic non-increasing trend in the first-dimension translation and scaling (but not in the second dimension transformations). That trend makes the template smaller and shifts it to the left as the sequence progresses. This drift can be seen in Figure 6: the peaks in the third standards chromatogram are left of the locations recorded in the template for the first standards run. The cumulative effect of this retention-time drift is illustrated in the last row of Table 1, which shows the transformation for matching the peaks of the sixth standards chromatogram (the last of the 64 runs) with the template from the first standards chromatogram (the first of the 64 runs).

Template matching can deal with retention-time drift in several ways. One way is to update the template throughout the sequence of runs as each new sequence is acquired. This approach yields excellent results, as suggested by Table 1, in which the transformation between any adjacent pair is relatively small. A consensus template can be built from the average of several recent datasets and updated to provide a "moving average" template. If there is substantial drift and no intermediate results with which to update the template, it may be necessary to increase the limits on the transformation space. Affine transformations have been shown to be adequate for modeling chromatographic drift over a large range of chromatographic conditions [25], but large nonlinear retention-time deformations may require more complex template transformations for peak matching.

Of course, the pattern matching problem in Figure 6 is simple: there are not many peaks, every peak in the template is detected in the chromatogram, and there are few other peaks in the chromatogram which might interfere with pattern recognition. In general, template matching works better and is more robust with more peaks because the matching is based on more data and is less susceptible to a few missing peaks or extra peaks. Of course, matching also is better if there is good separation of peaks—ideally, only one peak in each retention-time window. As peaks become less well-separated, template matching is more challenging, but as long as the pattern is maintained (i.e., peaks are detected in the same positions relative to each other, subject to the transformation) template matching is robust. Even overlapping peaks are not a problem as long as the pattern of detected peaks is maintained. However, template matching, like any identification method based on retention time, is subject to errors if new (unexpected) peaks that change the pattern are detected within the retention-time windows of peaks in the pattern, especially if the target peaks are not present. For these more difficult problems, template matching on only the chromatographic pattern (i.e., peak retention times) may not be sufficient to correctly identify all peaks of interest. The last example of this section presents data for which there are template matching errors, setting the stage for Smart Templates that augment templates with multispectral constraints (as described in the next section).

A more challenging problem is presented in Figure 7, which shows a LC × LC chromatogram of human urine, one of 14 control samples interspersed in the sequence of 64 samples. (A different color map is used to illustrate this example.) By visual examination of the chromatographic peaks detected in the control sample data, a set of 66 peaks was selected. Then, the template from each chromatogram was composed of the peaks from that set which were detected in the chromatogram. For example, peak detection for the chromatogram of the first control sample yielded 64 of the 66 peaks in the peak set, so the template generated from it contained those 64 peaks. As was done for the standards samples, the template from each control sample was matched to the peaks detected in the chromatogram of the next control sample. For example, when the template from the first control sample was matched to the chromatogram from the second control sample, 62 of the 64 peaks in the template were matched correctly.

The results for template matching with the control samples are summarized in Table 2. (The example in Figure 7 is in the third row.) A few explanations are required. First, if a peak was split during detection (i.e., incorrectly detected as two or more peaks) and if the template matched one of the parts of the split peak, the match was considered correct (with the logic that the match was to the correct peak). The example of Figure 7 was selected because it shows both types of *Peak Errors* and both types of *Match Errors*. Two types of problems were recorded as Peak Errors: (1) if no peak was detected, then the template could not match that peak, and (2) if two peaks were merged in detection, then the template matching could not match both peaks. The first type of peak error is noted by Arrow 1 in Figure 7 and the second type of peak error is indicated by Arrows 2, 3, and 4. Two types of errors were recorded as Match Errors: (1) if the peak was detected, but template matching did not match, and (2) if a peak was not detected (e.g., merged with another peak), but the template matched an incorrect peak. The first type of match error is indicated by Arrow 5 and the second type of match error is indicated by Arrows 6 and 7.

The success rate for template matching was high—97% overall. In that sense, Figure 7 is somewhat misleading because, among the 13 matched chromatograms, it accounted for 4 of the 19 peak errors and 3 of the 6 matching errors. Overall, 778 of the 803 peaks in the 13 templates were matched correctly in the next chromatogram. Of the 25 matching failures, 19 were peak errors, for which matching cannot succeed. There were only six match errors, an error rate of less than 1%.

The template matching parameters can be changed to eliminate some matching errors. For example, the matching error indicated by Arrow 5 can be eliminated by increasing the retention-time window within which peaks may be matched. Similarly, the matching errors indicated by Arrows 6 and 7 can be eliminated by reducing the retention-time window within which peaks may be matched. However, the tension between these two actions is problematic: which windows should be made smaller and which windows should be made bigger? The answer depends on the detected peaks, which are not known when the template is created. A better solution is to use additional logic in the templates, i.e., Smart Templates, as described next.

6. Smart Templates

Smart Templates use peak-specific constraints, such as multispectral matching, to reduce or eliminate template pattern-matching errors. The constraints are expressed in the Computer Language for Identifying Chemicals (CLIC) [28], augmented with the seven multispectral matching metrics introduced in Section 4. (CLIC is described more fully in Reference [28].) Each peak in a Smart Template can have a constraint rule, involving the spectrum of the peak (either at the apex or integrated over all data points in the peak), statistics about the peak (e.g., its fractional response as a part of the whole sample), and/or its retention time, combined with arithmetic, relational, and logical operators. For example, if the chemical identity of a peak is known and its expected spectrum is cataloged in a library, then matching for that peak can be restricted to peaks with sufficiently high multispectral match factor (or sufficiently low multispectral difference). In the example of this section, the rules constrain the Euclidean distance between the expected spectra in the Smart Template and the observed spectra in the data.

The constraints can provide greater selectivity during template matching, allowing two types of improvements. First, peaks which are within the retention-time window but which are not correct matches can be rejected. This improvement can eliminate the matching errors indicated by Arrows 6 and 7 in

Table 2. Results for template matching with the control urine samples

Template sequence #	Target sequence #	Template size	Number correct	Success rate (%)	Peak detection errors	Peak detection error rate (%)	Match errors	Match error rate (%)	Smart match errors	Smart match error rate (%)
3	7	64	62	97	2	3.1	0	0.0	0	0.0
7	11	62	60	97	2	3.2	0	0.0	0	0.0
11	15	61	54	89	4	6.6	3	4.9	0	0.0
15	19	58	57	98	1	1.7	0	0.0	0	0.0
19	21	58	57	98	0	0.0	1	1.7	0	0.0
21	25	62	62	100	0	0.0	0	0.0	0	0.0
25	29	64	62	97	1	1.6	1	1.6	0	0.0
29	33	64	61	95	3	4.7	0	0.0	0	0.0
33	37	62	62	100	0	0.0	0	0.0	0	0.0
37	39	63	62	98	1	1.6	0	0.0	0	0.0
39	43	62	60	97	2	3.2	0	0.0	0	0.0
43	47	62	61	98	1	1.6	0	0.0	0	0.0
47	63	61	58	95	2	3.3	1	1.6	0	0.0
Total		803	778	97	19	2.4	6	0.7	0	0.0

Figure 7 because the spectra of those peaks do not match the template spectra. Second, pursuant to the first improvement, the size of the retention-time matching window may be increased to allow more distant matches without increasing the number of incorrect matches allowed by the larger window if the constraint in the Smart Template rejects those incorrect matches. This improvement can eliminate the matching error indicated by Arrow 7 in Figure 7.

Some care is required in writing constraints for Smart Templates. For example, consider a constraint which requires that the Euclidean distance between the expected UV absorbance spectrum recorded in a template and the spectrum of a matched peak in the chromatogram be less than a specific value, expressed in CLIC as:

$$\text{Euclidean Distance}(\text{“<ms>”}) < 0.22 \qquad (1)$$

where “<ms>” refers to the expected multispectrum of the template peak (a mass spectrum or in this case a UV absorbance spectrum, which is recorded from the chromatogram(s) from which the template is created) and the spectrum of the peak considered for matching is implicit in the expression. (Both spectra are range normalized before computing the Euclidean distance.) With this rule, in the example of Figure 7, the matching errors indicated for Arrows 6 and 7 are eliminated (as is a similar error in the matching of the template for the thirteenth control sample to the chromatogram of the 14th). Those chromatographic peaks could be matched to the peak pattern based on the retention-time pattern, but their spectra are not similar enough to the target spectra recorded in the template.

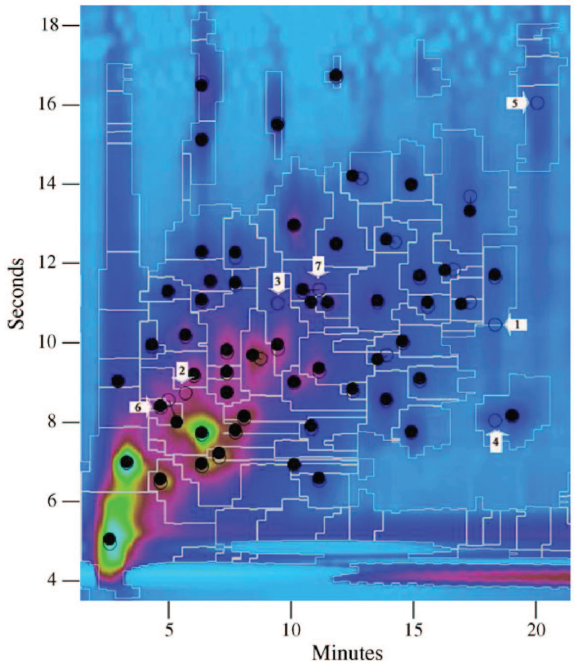

Figure 8. Smart Template matching for the fourth control sample with the template from the third control sample (same pair as Figure 7).

Note that such constraints might be so restrictive that other, correct matches are disallowed. For this chromatogram, Constraint (1) does not prevent correct matches of those three peaks in any of the 13 matchings. However, if used for all peaks in all matchings, that constraint will prevent correct matches in one or more matchings for the four top-rightmost peaks and one of the bottom-rightmost peaks, all of which are faint and so have lower signal-to-noise ratios. For those peaks, a different constraint threshold value is required. So, different values in the constraint (i.e., the threshold for multispectral difference) should be used for different peaks.

Automated constraint-building uses evaluations of the multispectral variability within the set of peaks for the same compound in one or more chromatograms and the multispectral differences with the set of peaks for other compounds. So, for example, if the spectral difference measured by Euclidean distance for peaks of the same compound is at most 0.1 and the spectral difference for peaks of other compounds is always greater than 0.3, then the automatically generated spectral rule requires a spectral difference of no more than the mid-point between the distances, 0.2 for this example. If only one chromatogram is used to construct the constraint, the maximum distance between peaks of the same compound is 0. The algorithm also is configurable to set a minimum and maximum distance used in the rule, so if the computed value is outside the user-defined range, it is thresholded. In cases that the spectral distance between two peaks for the same compound in two different chromatograms is larger than the spectral distance with a peak of another compound, the automated template building algorithm constructs the rule to always match correct compounds (even if some incorrect matches are allowed). So, for example, if the spectral distances for peaks of the same compound are as large as 0.1, then the value for the constraint would not be less than 0.1, even if the spectral distance for peaks of some other compounds is less than 0.1. (However, again this value is subject to a user-defined minimum and maximum value.) With this approach, all template peaks can be assigned constraints on Euclidean distance (or one of the other multispectral metrics) automatically.

These multispectral constraints eliminate all matching errors to incorrect peaks with the data presented in Figure 7 and with the other control sample chromatograms. As outlined above, the matching errors for peaks outside the retention-

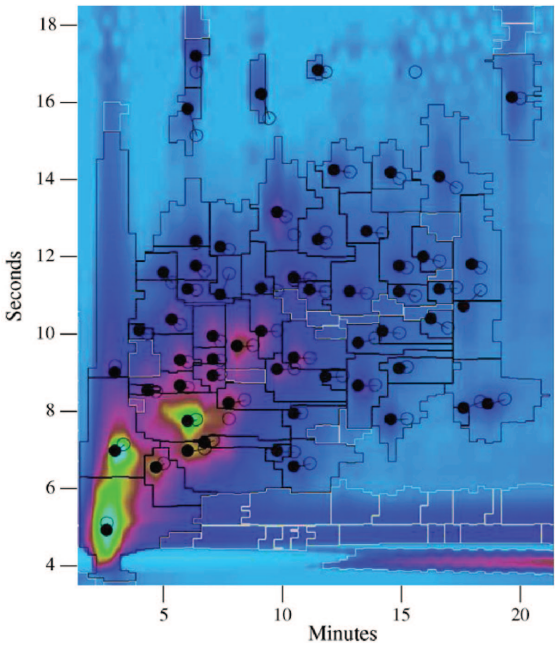

Figure 9. Smart Template matching for the template from the first control sample to the peaks of the 14th control sample. All matching errors are eliminated.

time matching window can be eliminated by increasing the size of the window for the secondary separation. With the multispectral constraints on the template peaks, increasing the window size does not cause any incorrect matches. So, with Smart Templates, constructed automatically, the matching error rate for the chromatograms of the control samples is reduced to zero, as shown in the last two columns of Table 2.

Figure 8 shows the results of Smart Template matching for the example chromatogram of Figure 7, with all matching errors (but not peak errors) eliminated. Figure 9 shows Smart Template matching of the template from the first of the control sample to the chromatographic peaks of the fourteenth control sample. As can be seen, the retention drift and template transformation for this example are greater. For template matching without constraints, 56 of 64 peaks were matched, with four peak errors and four matching errors. A Smart Template with constraints eliminates all matching errors.

7. Conclusion

With improved chromatographic performance, LC × LC is emerging as a powerful technology for complex separations, e.g., biochemical assays for proteomics and metabolomics [1]. Recent surveys of LC × LC research and development cite the lack of efficient and effective software as a significant impediment to fully realizing the benefits of these technological improvements [5, 6]. LC × LC transforms chemical samples into raw data; but advances in information technologies are required to transform complex LC × LC data into useful information.

This paper addresses the important problem of automatically identifying and classifying peaks, even with variable chromatographic conditions. Smart Templates record a peak pattern in a template with analytical metadata and constraints on peak identification. The template pattern is matched to find the similar pattern of peaks in target chromatograms, subject to the constraints and user-defined parameters. Then, the analytical metadata is copied onto the new data, thereby identifying and classifying peaks. With a transformation model flexible enough to account for chromatographic variations and selectively discriminating constraints, the approach is highly robust. In experiments analyzing 13 urine samples with 803 target analyte peaks, template matching on retention time only resulted six identification errors (0.7% error rate) and Smart Templates resulted in zero identification errors (0.0% error rate).

This powerful approach is demonstrated for a series of LC × LC separations of human urine with a UV detector, but the method is applicable to other multidimensional chemical separations such as GC × GC, HPLC with capillary electrophoresis (LC–CE), etc., and to other detectors, including mass spectrometers (which provide better sensitivity and selectivity for even more reliable peak matching). Smart Templates can be used to quickly and accurately match large numbers of peaks in complex patterns and so provide a powerful tool for LC × LC analyses.

Acknowledgments

The research described was supported by Grant Number 0450540 from the National Science Foundation and Grant Number 5R44RR20256 from National Center for Research Resources of the National Institutes of Health (NIH) and by a fellowship from the American Chemical Society to D. R. Stoll. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NSF nor the NIH.

References

[1] In: S. A. Cohen and M. R. Schure, Editors, *Multidimensional Liquid Chromatography: Theory and Applications in Industrial Chemistry and the Life Sciences*, John Wiley and Sons, New York, NY (2008).
[2] F. Erni and R. Frei, *J. Chromatogr.* **149** (1978), p. 561.
[3] M. Bushey and J. Jorgenson, *Anal. Chem.* **62** (2) (1990), pp. 161–167.
[4] R. Shellie and P. Haddad, *Anal. Bioanal. Chem.* **386** (3) (2006), p. 405.
[5] D. R. Stoll, X. Li, X. Wang, P. W. Carr, S. E. Porter, and S. C. Rutan, *J. Chromatogr. A* **1168** (2007), p. 3.
[6] G. Guiochon, N. Marchetti, K. Mriziq, and R. Shalliker, *J. Chromatogr. A* **1189** (1–2) (2008), p. 109.
[7] C. Evans and J. Jorgenson, *Anal. Bioanal. Chem.* **378** (8) (2004), p. 952.
[8] H. J. Issaq, *Electrophoresis* **22** (17) (2001), p. 3629.
[9] H. J. Issaq, K. C. Chan, T. P. Janini, M. George, T. D. Conrads, and Veenstra, *J. Chromatogr. B* **817** (1) (2005), p. 35.
[10] W. B. Dunn, N. J. C. Bailey, and H. E. Johnson, *Analyst* **130** (5) (2005), p. 606.
[11] D. I. E. Warwick and B. Dunn, *Trends Anal. Chem.* **24** (4) (2005), p. 285.
[12] M. Wagner, D. N. Naik, A. Pothen, S. Kasukurti, R. R. Devineni, B-L. Adam, O. J. Semmes, and G. L. Wright Jr., *BMC Bioinform.* **5** (26) (2004), p. 1.
[13] D. Radulovic, S. Jelveh, S. Ryu, T. G. Hamilton, E. Foss, Y. Mao, and A. Emili, *Mol. Cell. Proteomics* **3** (10) (2004), p. 984.
[14] A. Vlahou and M. Fountoulakisa, *J. Chromatogr. B* **814** (1) (2005), p. 11.
[15] P. J. Ulintz, J. Zhu, Z. S. Qin, and P. C. Andrews, *Mol. Cell. Proteomics* **5** (3) (2006), p. 497.
[16] S. P. Dixon, I. D. Pitfield, and D. Perrett, *Biomed. Chromatogr.* **20** (6–7) (2006), p. 508.
[17] D. Stoll, J. Cohen, and P. Carr, *J. Chromatogr. A* **1122** (1–2) (2006), p. 123.
[18] GC Image, LLC, GC Image®software, http://www.gcimage.com (2008).
[19] A. Visvanathan, S. E. Reichenbach, and Q. Tao, *J. Electron. Imaging* **16** (3) (2007), p. 033004.
[20] S. E. Reichenbach, M. Ni, D. Zhang, and E. B. Ledford Jr., *J. Chromatogr. A* **985** (1) (2003), p. 47.
[21] S. E. Reichenbach, M. Ni, V. Kottapalli, and A. Visvanathan, *Chemom. Intell. Lab. Syst.* **71** (2) (2004), p. 107.
[22] S. Beucher, C. Lantuejoul, in: International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation, 1979, p. 17.
[23] A. E. Sinha, C. G. Fraga, B. J. Prazen, and R. E. Synovec, *J. Chromatogr. A* **1027** (1–2) (2004), p. 269.
[24] S. Porter, D. Stoll, S. Rutan, P. Carr, and J. Cohen, *Anal. Chem.* **78** (15) (2006), p. 5559.
[25] M. Ni, S. E. Reichenbach, A. Visvanathan, J. R. TerMaat, and E. B. Ledford Jr., *J. Chromatogr. A* **1086** (1–2) (2005), p. 165.
[26] M. Ni, S. E. Reichenbach, in: Proceedings of the International Conference on Pattern Recognition, vol. 2, IAPR/IEEECS, 2004, p. 145.
[27] M. Ni, Point Pattern Matching and its Application in GC × GC, Ph.D. thesis, University of Nebraska (2004).
[28] S. E. Reichenbach, V. Kottapalli, M. Ni, and A. Visvanathan, *J. Chromatogr. A* **1071** (1–2) (2004), p. 263.